大语言模型旋转位置编码的简易推导

何沧平
cangping@staff.weibo.com
微博

许涛
xutao@sugon.com
曙光信息产业（北京）有限公司

摘　要

以 LLAMA 为代表的开源大语言模型广泛使用旋转位置编码，原始论文使用复函数推导。本文改用线性代数推导，期望更好地理解该编码方法；提出该方法的一个疑点并给出了改进建议。

关键词: 大语言模型, LLM, 旋转位置编码, LLAMA

Easy Derivation Of Rotary Position Embeddings For Large Language Models*

He Cangping
cangping@staff.weibo.com
WEIBO.COM

Xu Tao
xutao@sugon.com
SUGON.COM

Abstract

The Rotary Position Embeddings(RoPE) is widely used in open-source large language models such as LLAMA. In the original paper, the formula derivation uses complex functions. In this Paper, I derive PoPE's formulas again with linear algebra, hoping to better understand this method.

Keywords: Large Language Model(LLM), Rotary Position Embeddings(RoPE), LLAMA

1　引言

2022 年 11 月发布的 ChatGPT 引爆新一轮科技创新，LLAMA 模型 [2] 开源后，一大批大语言模型发布，例如 Baichuan-7B [1] 采用了与 LLAMA 相同的模型结构。旋转位置编码 (Rotary Position Embeddings, RoPE)[1] 是 LLAMA 模型的一个重要组件，原始论文使用复变函数的理论来推导，对不熟悉复变函数的人来说不容易理解。本文尝试使用更常用见的线性代数来推导，期望能让更多人理解这个优秀的编码方法。

*完稿日期：2023 年 7 月 10 日

[1] https://github.com/baichuan-inc/baichuan-7B

2 函数定义

作为准备，本节定义几个函数。目前 pytorch 代码中数组的组织方式是行优先，序号从 0 开始，因此本文中的向量、矩阵也按行优先来定义，矩阵元素的序号也从 0 开始。

任意给定正整数 m 和 n，行向量用黑体小写字母表示，形式为 $\boldsymbol{x} = (x_0, x_1, \ldots, x_{n-1})$。矩阵用大写字母表示，形式为

$$X = \begin{bmatrix} x_{00} & x_{01} & \cdots & x_{0,n-1} \\ x_{10} & x_{11} & \cdots & x_{1,n-1} \\ \vdots & \vdots & & \vdots \\ x_{m-1,0} & x_{m-1,1} & \cdots & x_{m-1,n-1} \end{bmatrix}.$$

软大函数 (softmax) 定义为

$$\begin{aligned} \mathrm{smax}(\boldsymbol{x}) &= \frac{1}{\sum_{i=0}^{i=n-1} e^{x_i}}(e^{x_0}, e^{x_1}, \ldots, e^{x_{n-1}}), \\ \mathrm{smax}(X) &= \begin{bmatrix} \mathrm{smax}(x_{0:}) \\ \mathrm{smax}(x_{1:}) \\ \vdots \\ \mathrm{smax}(x_{m-1:}) \end{bmatrix} = (\mathrm{smax}(x_{0:}); \mathrm{smax}(x_{1:}); \ldots; \mathrm{smax}(x_{m-1:})), \end{aligned}$$

这里的 $x_{i:} = (x_{i0}, x_{i1}, \ldots, x_{i,n-1})$，圆括号里的分号表示换行。

3 旋转位置编码

在 LLAMA 模型中，自注意力的核心运算是软大函数，即用给定“查”矩阵 Q 和“值”矩阵 K，计算

$$\mathrm{smax}\left(\frac{QK^T}{\sqrt{n_6}}\right),$$

这里 Q 的尺寸是 $n_3 \times n_6$，n_3 是当前序列长度，每生成一个词碎（token）后加 1；n_6 是单个注意力头的宽度，在 LLAMA-7B 中 $n_6 = 128$。K 的尺寸与 Q 的尺寸相同，元素值不同。记 $R = QK^T$，矩阵 R 的尺寸是 $n_3 \times n_3$。将矩阵 Q、K、R 的形式分别记为

$$Q = \begin{bmatrix} q_{0:} \\ q_{1:} \\ \vdots \\ q_{n_3-1:} \end{bmatrix}, \quad K = \begin{bmatrix} k_{0:} \\ k_{1:} \\ \vdots \\ k_{n_3-1:} \end{bmatrix}, \quad R = \begin{bmatrix} r_{00} & r_{01} & \cdots & r_{0,n_3-1} \\ r_{10} & r_{11} & \cdots & r_{1,n_3-1} \\ \vdots & \vdots & & \vdots \\ r_{n_3-1,0} & r_{n_3-1,1} & \cdots & r_{n_3-1,n_3-1} \end{bmatrix}.$$

向量 $q_{i:}$ 和 $k_{i:}$ 对应当前序列里的第 i 个词碎，$i = 0, 1, \cdots, n_3$，实数 r_{ij} 是 $q_{i:}$ 和 $k_{j:}$ 的内积，即

$$r_{ij} = q_{i:} k_{i:}^T \tag{1}$$

位置编码的目标是对向量 $q_{i:}$ 和 $k_{j:}$ 分别施加一个带绝对位置信息 i 和 j 的变换 f，即 $\hat{r}_{ij} = f(q_{i:}) f(k_{i:})^T$，以满足下列要求：

(c1) 在 $i = j$ 时，施加变换后内积保持不变，即 $\hat{r}_{ii} = r_{ii}$。对应到词碎序列上的意义是，任意位置上词碎的自身内积不受变换 f 影响。

(c2) 变换后它们的内积 $\hat{r}_{ij}$ 只包含相对位置信息 $i - j$，不再包含绝对位置信息。

(c3) 对元素值固定的 $q_{i:}$ 和 $k_{j:}$，$|i-j|$ 越大，$|\hat{r}_{ij}-r_{ij}|$ 越小。对应到词碎序列上的意义是，两词碎离得越远，相互影响越小。

作为科研的基本套路，先看最简单情形，$n_6=2$。向量分别写出来，$q_{i:}=(q_{i0},q_{i1})^T$, $k_{j:}=(k_{j0},k_{j1})^T$。假设变换 f 是一个线性变换，即 $f(q_{i:})=q_{i:}A_i$，这里的 A_i 是一个尺寸为 2×2 的矩阵。从而有

$$\hat{r}_{ij}=f(q_{i:})f(k_{i:})^T=q_{i:}A_i(k_{j:}A_j)^T=q_{i:}A_iA_j^Tk_{j:}^T \tag{2}$$

对式 (2) 应用 (c1)，可以得到

$$A_iA_i^T=I,\quad 且A_i\neq I. \tag{3}$$

式 (3) 要求 A_i 是正交矩阵，而线性代数中常见的正交矩阵是旋转矩阵

$$I_\theta=\begin{bmatrix}\cos\theta & \sin\theta\\ -\sin\theta & \cos\theta\end{bmatrix},$$

容易验证 $I_\theta I_\theta^T=I$。

注意到，$q_{i:}I_\theta$ 的几何含义是将向量 $q_{i:}$ 旋转弧度 θ，$q_{i:}I_\theta I_\theta^T$ 的几何含义是将向量 $q_{i:}$ 旋转弧度 θ 后，再旋转弧度 $-\theta$，两次旋转的弧度抵消，向量值保持不变。如果两次旋转的弧度不同，那最终效果就是旋转一个差值弧度。因此，记位置 i 和 j 上的旋转弧度分别是 θ_i 和 θ_j，容易验证

$$I_{\theta_i}I_{\theta_j}^T=\begin{bmatrix}\cos\theta_i & \sin\theta_i\\ -\sin\theta_i & \cos\theta_i\end{bmatrix}\begin{bmatrix}\cos\theta_j & -\sin\theta_j\\ \sin\theta_j & \cos\theta_j\end{bmatrix}=\begin{bmatrix}\cos(\theta_i-\theta_j) & \sin(\theta_i-\theta_j)\\ -\sin(\theta_i-\theta_j) & \cos(\theta_i-\theta_j)\end{bmatrix}$$

记 $\theta_{ij}=\theta_i-\theta_j$，将 $A_i=I_{\theta_i}$ 和 $A_j=I_{\theta_j}$ 代入 (2)，得到

$$\hat{r}_{ij}=q_{i:}(I_{\theta_i}I_{\theta_j}^T)k_{j:}^T=q_{i:}I_{\theta_{ij}}k_{j:}^T. \tag{4}$$

显然，式 (4) 只包含相对位置信息 θ_{ij}，满足要求 (c2)。

词碎序列中的每个位置 $i=0,1,\dots,n_3-1$，都要对应一个 θ_i。原始论文 [1] 的选择是一个等差数列，即先选定 $\bar{\theta}_0$，然后令 $\theta_i=i\bar{\theta}_0$。

当 n_6 是大于 2 的偶数时，可以二维一组地处理，即对长度为 n_6 的行向量 $q_{i:}=(q_{i0},q_{i1},\dots,q_{i,n_6-1})$ 作旋转变换 $q_{i:}A_i$，

$$A_i=\begin{bmatrix}\cos i\bar{\theta}_0 & \sin i\bar{\theta}_0 & 0 & 0 & \cdots & 0 & 0\\ -\sin i\bar{\theta}_0 & \cos i\bar{\theta}_0 & 0 & 0 & \cdots & 0 & 0\\ 0 & 0 & \cos i\bar{\theta}_2 & \sin i\bar{\theta}_2 & \cdots & 0 & 0\\ 0 & 0 & -\sin i\bar{\theta}_2 & \cos i\bar{\theta}_2 & \cdots & 0 & 0\\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & 0 & 0 & \cdots & \cos i\bar{\theta}_{n_6-2} & \sin i\bar{\theta}_{n_6-2}\\ 0 & 0 & 0 & 0 & \cdots & -\sin i\bar{\theta}_{n_6-2} & \cos i\bar{\theta}_{n_6-2}\end{bmatrix}$$

对于弧度 $\bar{\theta}_t$，$t=0,2,6,\dots,n_6-2$，原始论文 [1] 使用固定值

$$\bar{\theta}_t=10000^{-t/n_6}. \tag{5}$$

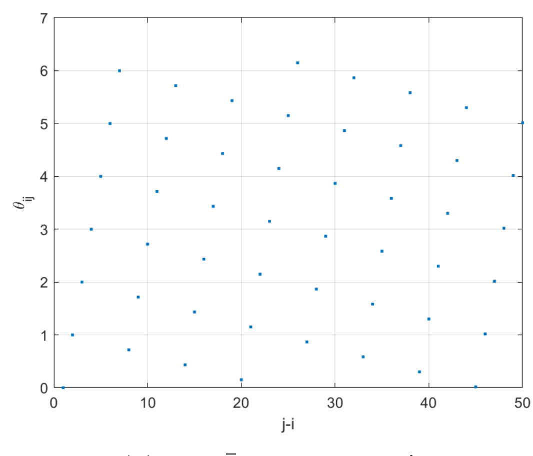


图 1: 当 $\bar{\theta}_0 = 1$ 时，内积中的旋转弧度

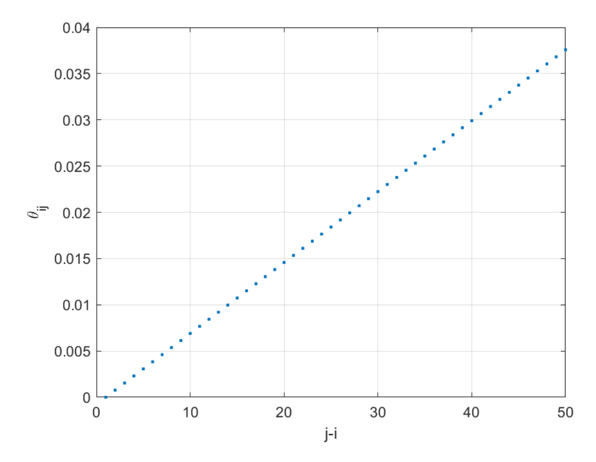


图 2: 当 $\bar{\theta}_0 = \frac{\pi}{2\times 2048}$ 时，内积中的旋转弧度

4 疑点与建议

按照式 (4) 计算，$\bar{\theta}_0 = 1$，对词碎位置差值 $j-i = 1, 2, \ldots, 50$ 和向量的前 2 维分量，内积运算式 (4) 中的弧度差是 $\theta_{ij} = (j-i)\bar{\theta}_1$，模 2π 后的变化走势见图1，当 $j-i = 6$ 时，弧度差 $\theta_{ij} = 6$；当 $j-i = 7$ 时，弧度差 $\theta_{ij} = 0.7168$。相距更远的两个词碎，旋转弧度差反而更小，不符合要求 (c3)。

本文建议尝试将式 (4) 更改为

$$\bar{\theta}_t = \frac{\pi}{2n_9} 10000^{-t/n_6},$$

这里的 n_9 是序列的最大长度，例如取值 2048。更改之后的效果见图2，满足要求 (c3)。

参考文献

[1] Jianlin Su. *RoFormer: Transformer with Rotary Position Embeddings - ZhuiyiAI*. Tech. rep. 2021. URL: https://github.com/ZhuiyiTechnology/roformer.

[2] Hugo Touvron et al. "LLaMA: Open and Efficient Foundation Language Models". In: (2023). arXiv: 2302.13971 [cs.CL].